

# GenePattern:

A platform for integrative genomics

Michael Reich

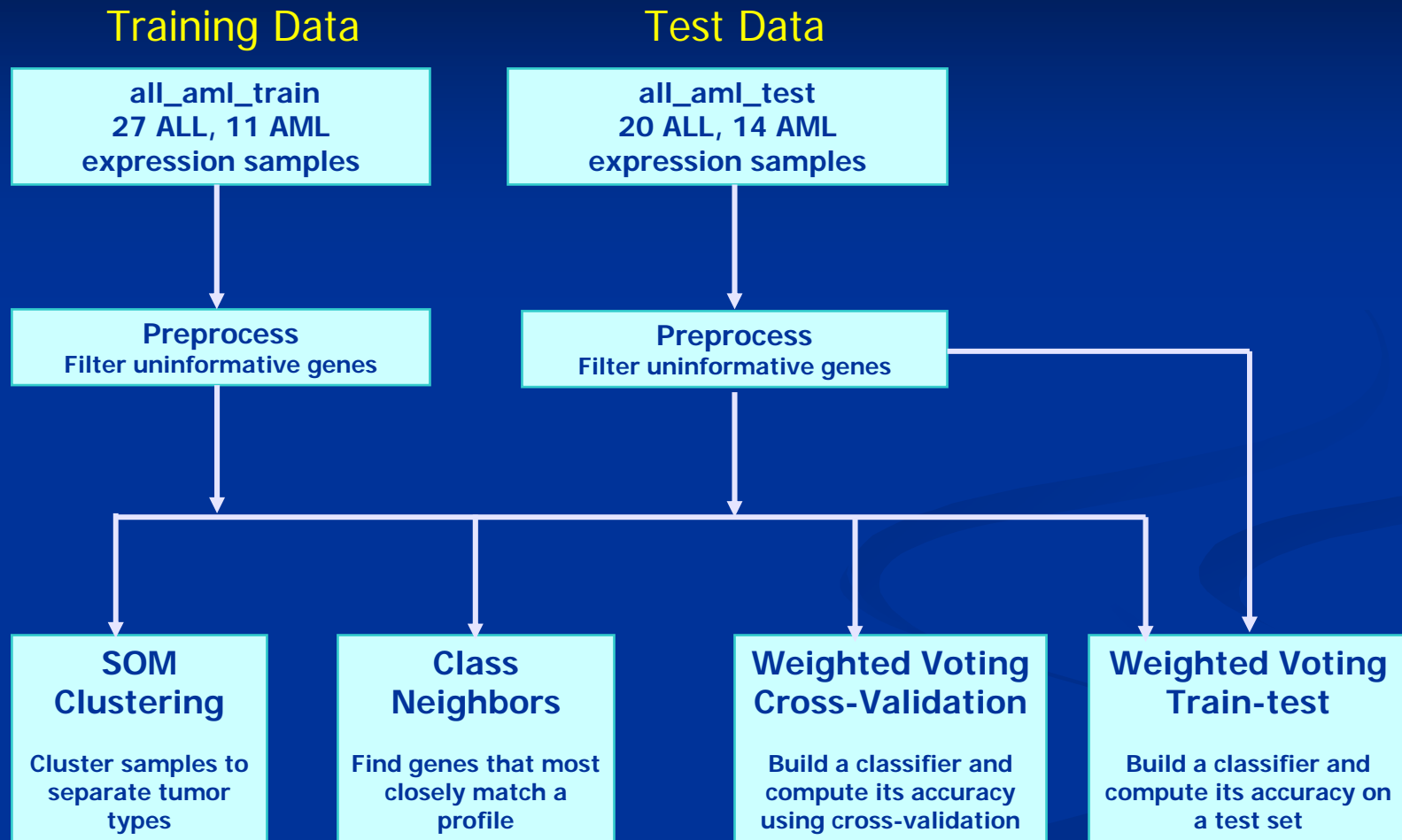
Broad Institute of MIT and Harvard

July 19, 2006

# Challenges in genomic research

- The number of available research tools is growing exponentially.
- Published results of *in silico* research are not reproducible without extensive communication with the authors.
- Research teams comprise members from many disciplines and levels of computational sophistication.
- The research environment is dynamic and heterogeneous, with new tools being developed quickly in many different forms.

# GenePattern initial DBP: ALL/AML Classification



Golub and Slonim et al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression*, Science 1999

# Design Objectives

- A comprehensive repository of analysis modules
- User interface accessible to users at all levels of computational sophistication
- Ability to chain tasks into pipelines for reproducible *in silico* research
- Ability to add new tools without programming
- Computing on a local machine or distributed among more powerful compute servers

# GenePattern: A platform for integrative genomics

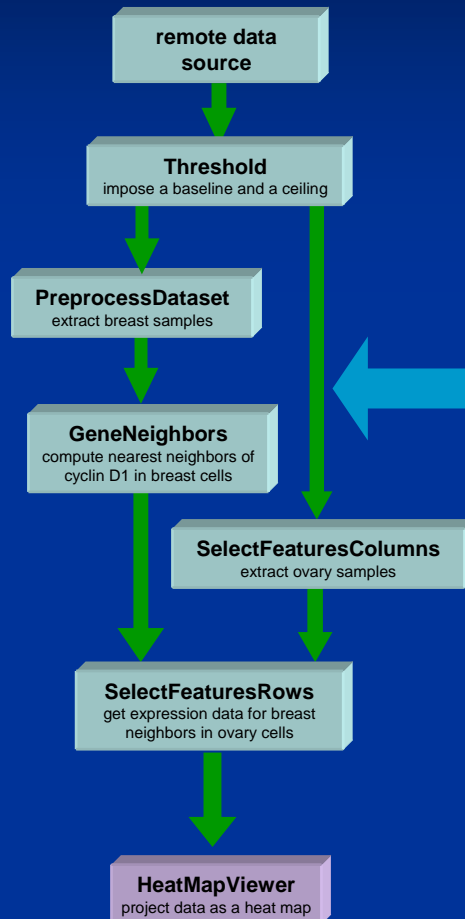
## Module Repository



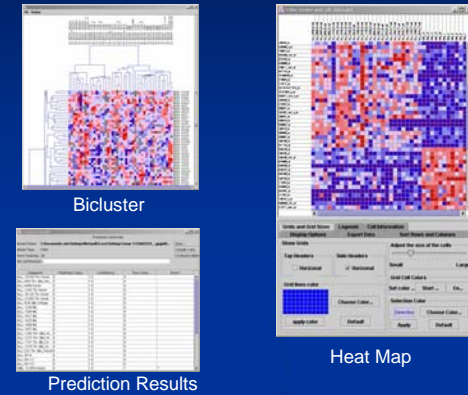
## Task Integrator

The screenshot shows the "add GenePattern task" window. It includes fields for Name, Description, Author, Owner, Privacy, Quality level, Command line, Task type, CPU type, Operating system, Language, and Module version. A "Support files" section at the bottom states: "The actual program plus any required libraries will be accessible to your command line".

## Pipeline Environment



## Graphical Environment



## Programming Environment

```
# source("D:/GSP2003/GenePattern_modules/Golub_et_al_1999.R", echo = TRUE)
# GenePattern
# Molecular Classification of Cancer: Class Prediction by Gene Expression
# Summary: This R/GenePattern script implements the supervised prediction method
# in Golub et al 1999. Science 286:531-537 (1999).

# Load and set up GenePattern commands and server
source("http://wilkins.mit.edu/7070/gsp/GenePattern.R", echo = FALSE, print.env
server <- SOAPServer("http://wilkins.mit.edu", "axis/serve/AxisServlet", 7
source(paste("http://", server@host, ".server@port", "gspGetAllTaskWrappers.j
```

```
MS.out <- MarkerSelection("data.filename" = "http://www.genome.mil.edu/mpr/pu
"data.results.file" = "pred.results",
"data.results.file" = "data.results",
"num.permutations" = "25")

file.show(MS.out$pred.results)
file.show(MS.out$data.results.gct)

data <- read.table(MS.out$pred.results, header=T, sep="T", skip=14)
```

## GenePattern

### Analysis Task Manager



# GenePattern Features

## Comprehensive Module Repository

- ~70 modules: analysis, visualization, pipelines
- Expression, proteomic, sequence, *variation (SNP)*, and *whole genome association* data
- Task suites

## Analytic Reproducibility

- Easy, rapid sharing of methodologies via pipelines
- Versioning using Life Sciences Identifier (LSID)
- Executable history of all sessions
- Automatic pipeline generation from result files
- Executable research documents

## Automatic Task Integration

- Add new modules without writing code
- Supports any command line callable code (language independent)

## Multiple user interfaces

- Java client
- Web client
- Programmatic interfaces to Java, MATLAB, R, *Perl*

## Local and Distributed Computing

- Laptop
- Client/Server
- Compute farm
- *Web Portal*

## Interoperability

- caBIG
  - Interface to caArray
  - *caGrid service (Beta)*
- *geWorkbench (Beta)*

# GenePattern Software

## Release Information

- Initially released in March, 2004
- Current version 2.0.2, released July 2006
- Currently ~3000 users, 500+ organizations, ~90 countries

## Availability

- Freely available
- Windows, Mac OS, and Linux platforms

## Resources

- <http://www.broad.mit.edu/genepattern>
- User workshops, documentation, email help desk, online user forum
- Reich et al. (2006) *Nature Genetics*



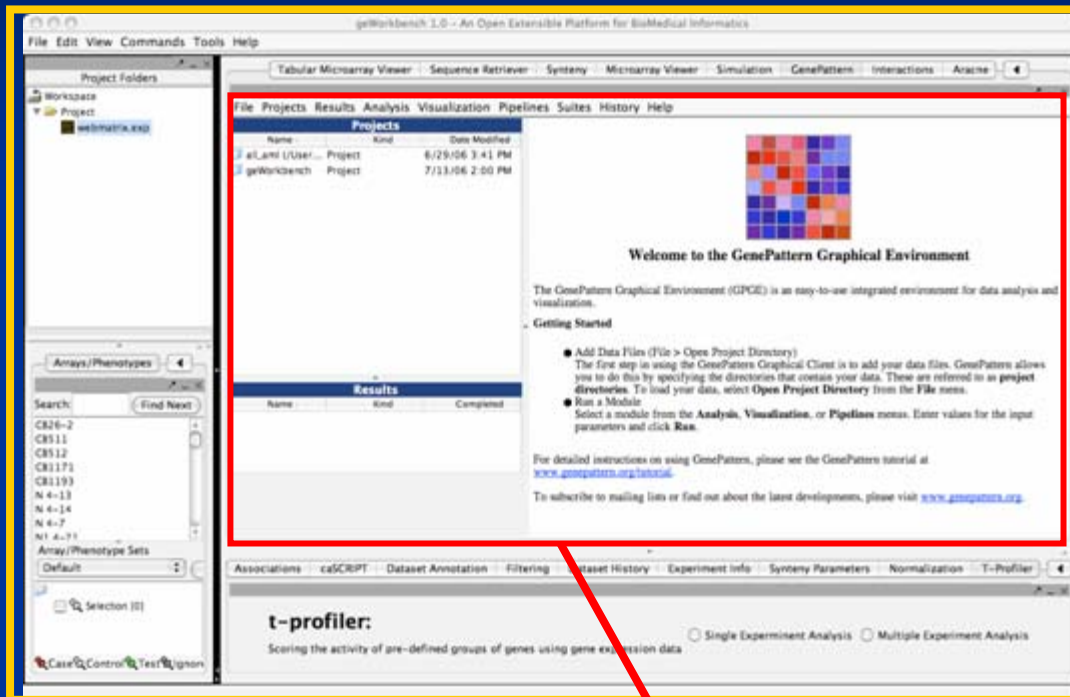
GenePattern is a winner of the  
2005 BioIT World Best Practices Award

# GenePattern/NCIBI Plan

- Integration with NCIBI software to support driving biological problems and bioinformatics tools
- Delivery of live and Web-based GenePattern training workshops



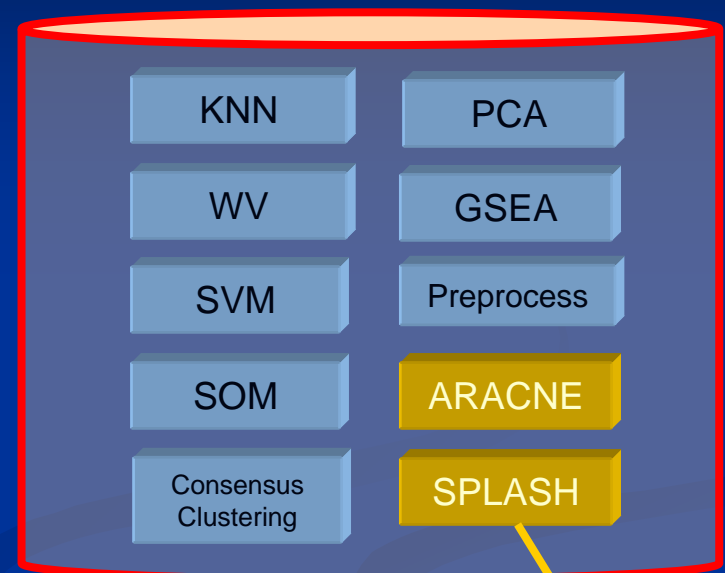
# GenePattern/geWorkbench Interoperability with MAGNet NCBC and NCIBI NCBC



geWorkbench  
application

GenePattern UI  
plug-in

Execute GenePattern modules from  
within geWorkbench



GenePattern  
module  
repository

geWorkbench  
modules

Wrap geWorkbench modules as  
GenePattern tasks

# Acknowledgements

## Core Development

Josh Gould  
Marc-Danie Nazaire  
Jim Lerner  
Ted Liefeld  
Jim Robinson  
David Twomey

## Module Contributors

Gad Getz  
Justin Lamb  
D.R. Mani  
Stefano Monti  
Ken Ross  
Aravind Subramanian

Todd R. Golub  
Pablo Tamayo  
Jill P. Mesirov, PI

[www.broad.mit.edu/genepattern](http://www.broad.mit.edu/genepattern)

